Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of Approximation Theory 164 (2012) 1513–1519

Journal of Approximation Theory

www.elsevier.com/locate/jat

Full length article

The convergence rate of a regularized ranking algorithm

Hong Chen

College of Science, Huazhong Agricultural University, Wuhan 430070, PR China

Received 20 March 2011; received in revised form 30 March 2012; accepted 4 September 2012
Available online 13 September 2012

Communicated by Ding-Xuan Zhou

Abstract

In this paper, we investigate the generalization performance of a regularized ranking algorithm in a reproducing kernel Hilbert space associated with least square ranking loss. An explicit expression for the solution via a sampling operator is derived and plays an important role in our analysis. Convergence analysis for learning a ranking function is provided, based on a novel capacity independent approach, which is stronger than for previous studies of the ranking problem.

Keywords: Ranking; Reproducing kernel Hilbert space; Sampling operator; Integral operator; Convergence rate

1. Introduction

Let us recall some basic concepts of statistical learning theory in the ranking setting (see [2] and references therein for details).

Let $\mathscr{X}$ be a compact metric space and $\mathscr{Y} = [0, M]$ for some $M > 0$. The relation between the input $x \in \mathscr{X}$ and the output $y \in \mathscr{Y}$ is described by a probability distribution $\rho(x, y) = \rho(y|x)\rho_{\mathscr{X}}(x)$ on $\mathscr{Z} := \mathscr{X} \times \mathscr{Y}$, where $\rho(y|x)$ is the conditional probability of $\mathscr{Y}$ given x and $\rho_{\mathscr{X}}(x)$ is the marginal probability of x. In learning theory, the distribution ρ is known only through a set of samples $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m \in \mathscr{Z}^m$ independently drawn according to ρ. Given samples $\mathbf{z}$, the ranking problem aims at finding a function $f_{\mathbf{z}} : \mathscr{X} \to \mathbb{R}$ that ranks future input instances with larger labels higher than those with smaller labels: x is to be ranked as preferred over x' if $y - y' > 0$, and lower than x' if $y - y' < 0$. In particular, $y - y' = 0$ indicates that there is no difference in ranking preference between the two input instances.

E-mail addresses: chenhongml@163.com, chenh@mail.hzau.edu.cn.

0021-9045/$ - see front matter
doi:10.1016/j.jat.2012.09.001

The quality of a ranking function f can be measured by its expected ranking error

$$\mathcal{E}(f) = \int_{\mathscr{Z}} \int_{\mathscr{Z}} (y - y' - (f(x) - f(x')))^2 d\rho(x, y) d\rho(x', y').$$

Denote by $\mathcal{F}$ the measurable function space and define $\mathcal{G} = \{f \in \mathcal{F} : f = \arg\min_{f \in \mathcal{F}} \mathcal{E}(f)\}$ as the target function set. We can observe that the target function is not unique and the regression function $f_\rho \in \mathcal{G}$, where

$$f_\rho(x) = \int_{\mathscr{Y}} y d\rho(y|x), \quad x \in \mathscr{X}.$$

It is easy to see that $\|f_\rho\|_\infty \le M$.

The ranking algorithm that we investigate in this article is based on a Tikhonov regularization scheme associated with a Mercer kernel—we usually call a symmetric and positive semidefinite continuous function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ a Mercer kernel. The reproducing kernel Hilbert space $\mathcal{H}_K$ associated with the kernel K is defined (see [3]) to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in \mathscr{X}\}$ with the inner product $\langle \cdot \rangle_K$ given by $\langle K_x, K_{x'} \rangle_K = K(x, x')$. The reproducing property takes the form $f(x) = \langle f, K_x \rangle_K, \forall x \in \mathscr{X}, f \in \mathcal{H}_K$. The reproducing property with the Schwartz inequality yields that $|f(x)| \le \sqrt{K(x, x)} \, \|f\|_{\mathcal{H}_K}$. Then $\|f\|_\infty \le \kappa \, \|f\|_{\mathcal{H}_K}$, where $\kappa := \sup_{x \in \mathscr{X}} \sqrt{K(x, x)}$.

The regularized ranking algorithm is implemented by an regularization scheme [2] in $\mathcal{H}_K$:

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m^2} \sum_{i,j=1}^{m} (y_i - y_j - (f(x_i) - f(x_j)))^2 + \lambda \|f\|_{\mathcal{H}_K}^2 \right\}, \tag{1}$$

where $\lambda > 0$ is the regularization parameter. A data free regularization function is

$$f_\lambda = \arg\min_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) + \lambda \|f\|_{\mathcal{H}_K}^2 \right\}. \tag{2}$$

Though the regularized ranking algorithm (1) has been well explained in [2] on the basis of stability analysis, it might be interesting and important to further investigate the convergence rate of $\inf_{f \in \mathcal{G}} \|f_{\mathbf{z},\lambda} - f\|_{\mathcal{H}_K}$. Though the convergence analysis for classification and regression algorithms has been elegantly studied in [9–11], there is no similar analysis in a ranking setting. The main difficulty here lies in the double-index summation in (1) and we shall tackle this problem by using a McDiarmid–Bernstein type of probability inequality for vector-valued random variables [6] with values in Hilbert spaces.

We mainly analyze the errors $\|f_{\mathbf{z},\lambda} - f_\lambda\|_{\mathcal{H}_K}$ and $\inf_{f \in \mathcal{G}} \|f_{\mathbf{z},\lambda} - f\|_{\mathcal{H}_K}$. There are three features for our convergence analysis. Firstly, the theoretical analysis provided here is capacity independent compared with error estimates under capacity assumptions in [1,4,7,8]; secondly, our capacity independent approach of convergence analysis is established from the viewpoint of operator approximation, unlike the stability analysis in [2,5]; finally, to the best of our knowledge, the proposed results for convergence in $\mathcal{H}_K$-norm are stronger than those in all the previous related studies.

2. Convergence analysis

Following the previous studies on Shannon sampling in [9,10], we define the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$ associated with a discrete subset $\mathbf{x} = \{x_i\}_{i=1}^m$ of $\mathscr{X}$ by

$$S_{\mathbf{x}}(f) = (f(x_i))_{i=1}^m = (f(x_1), \ldots, f(x_m))^T.$$

The adjoint of the sampling operator, $S_{\mathbf{x}}^T : \mathbb{R}^m \to \mathcal{H}_K$, is given by

$$S_{\mathbf{x}}^T c = \sum_{i=1}^m c_i K_{x_i}, \quad c = (c_i)_{i=1}^m = (c_1, \ldots, c_m)^T \in \mathbb{R}^m.$$

Define $D = m\mathbf{I} - \mathbf{1}\mathbf{1}^T$ and $Y = (y_i)_{i=1}^m = (y_1, \ldots, y_m)^T$, where $\mathbf{I}$ is the mth-order unit matrix and $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^m$.

Since $\langle S_{\mathbf{x}}^T D S_{\mathbf{x}}(f), f\rangle_K = \sum_{i,j=1}^m (f(x_i) - f(x_j))^2$, we can observe the following properties of operator $S_{\mathbf{x}}^T D S_{\mathbf{x}}$.

Lemma 1. *The operator $S_{\mathbf{x}}^T D S_{\mathbf{x}}$ is a self-adjoint positive linear operator.*

Using this notation and these properties, we can prove the following expression for $f_{\mathbf{z},\lambda}$.

Lemma 2. *The solution $f_{\mathbf{z},\lambda}$ defined in (1) can be expressed as*

$$f_{\mathbf{z},\lambda} = \left(\frac{1}{m^2} S_{\mathbf{x}}^T D S_{\mathbf{x}} + \frac{\lambda}{2} I\right)^{-1} \frac{1}{m^2} S_{\mathbf{x}}^T D Y. \tag{3}$$

Proof. Define $\mathcal{E}_{\mathbf{z}}(f) = \frac{1}{m^2} \sum_{i,j=1}^m (y_i - y_j - (f(x_i) - f(x_j)))^2$. By means of the reproducing property $f(x) = \langle f, K_x\rangle_K$, we know that

$$\begin{aligned}\frac{\partial(\mathcal{E}_{\mathbf{z}}(f) + \lambda\|f\|_{\mathcal{H}_K}^2)}{\partial f} &= \frac{4}{m^2} \sum_{i,j=1}^m (y_i K_{x_j} - y_i K_{x_i} + f(x_i)K_{x_i} - f(x_i)K_{x_j}) + 2\lambda f \\ &= 4\left(\frac{1}{m^2} S_{\mathbf{x}}^T D S_{\mathbf{x}} + \frac{\lambda}{2} I\right) f - \frac{4}{m^2} S_{\mathbf{x}}^T D Y.\end{aligned}$$

Then $f_{\mathbf{z},\lambda}$ is given by the solution $\frac{\partial(\mathcal{E}_{\mathbf{z}}(f)+\lambda\|f\|_{\mathcal{H}_K}^2)}{\partial f} = 0$ and has the expression

$$f_{\mathbf{z},\lambda} = \left(\frac{1}{m^2} S_{\mathbf{x}}^T D S_{\mathbf{x}} + \frac{\lambda}{2} I\right)^{-1} \frac{1}{m^2} S_{\mathbf{x}}^T D Y.$$

This proves our conclusion. $\square$

Define the integral operator $L_K : L^2_{\rho_{\mathscr{X}}} \to \mathcal{H}_K$ as

$$L_K f = \int_{\mathscr{X}} \int_{\mathscr{X}} f(x)(K_x - K_{x'}) d\rho_{\mathscr{X}}(x) d\rho_{\mathscr{X}}(x').$$

Note that

$$\langle L_K f, f\rangle_K = \int_{\mathscr{X}} f^2(x) d\rho_{\mathscr{X}}(x) - \left(\int_{\mathscr{X}} f(x) d\rho_{\mathscr{X}}(x)\right)^2 \ge 0.$$

The operator L_K can also be defined as a self-adjoint positive linear operator on $\mathcal{H}_K$. The regularization function defined in (2) satisfies $(L_K + \frac{\lambda}{2} I) f_\lambda = L_K f_\rho$. Then

$$f_\lambda = \left(L_K + \frac{\lambda}{2} I\right)^{-1} L_K f_\rho. \tag{4}$$

To bound the sample error $\|f_{\mathbf{z},\lambda} - f_\lambda\|_{\mathcal{H}_K}$, we introduce the following McDiarmid–Bernstein type of probability inequality for vector-valued random variables proved in [6].

Lemma 3. *Let $\mathbf{z} = \{z_i\}_{i=1}^m$ be independently drawn according to a probability distribution ρ on $\mathscr{Z}$, $(H, \|\cdot\|)$ be a Hilbert space, and $F : \mathscr{Z}^m \to H$ be measurable. If there is $\tilde{M} \geq 0$ such that $\|F(\mathbf{z}) - \mathbb{E}_{z_i} F(\mathbf{z})\| \leq \tilde{M}$ for each $1 \leq i \leq m$ and almost every $\mathbf{z} \in \mathscr{Z}^m$, then for every $\varepsilon > 0$,*

$$\mathrm{Prob}_{\mathbf{z} \in \mathscr{Z}^m} \{\|F(\mathbf{z}) - \mathbb{E}_{\mathbf{z}}(F(\mathbf{z}))\| \geq \varepsilon\} \leq 2 \exp\left\{-\frac{\varepsilon^2}{2(\tilde{M}\varepsilon + \sigma^2)}\right\},$$

where $\sigma^2 := \sum_{i=1}^m \sup_{\mathbf{z}\setminus\{z_i\} \in \mathscr{Z}^{m-1}} \mathbb{E}_{z_i}\{\|F(\mathbf{z}) - \mathbb{E}_{z_i}(F(\mathbf{z}))\|^2\}$. For any $0 < \delta < 1$, with confidence $1 - \delta$, it holds that

$$\|F(\mathbf{z}) - \mathbb{E}_{\mathbf{z}}(F(\mathbf{z}))\| \leq 2(\tilde{M} + \sqrt{\sigma^2}) \log \frac{2}{\delta}.$$

Proposition 4. *For any $0 < \delta < 1$, with confidence at least $1 - \delta$, we have*

$$\|f_{\mathbf{z},\lambda} - f_\lambda\|_{\mathcal{H}_K} \leq \frac{16 M\kappa(1 + \kappa\lambda^{-\frac{1}{2}})}{\lambda\sqrt{m}} \log \frac{2}{\delta} + \frac{M}{2m\sqrt{\lambda}}.$$

Proof. On the basis of (3) and (4), we have

$$f_{\mathbf{z},\lambda} - f_\lambda = \left(\frac{1}{m^2} S_{\mathbf{x}}^T D S_{\mathbf{x}} + \frac{\lambda}{2} I\right)^{-1} \left(\frac{1}{m^2} S_{\mathbf{x}}^T D Y - \frac{1}{m^2} S_{\mathbf{x}}^T D S_{\mathbf{x}} f_\lambda - \frac{\lambda}{2} f_\lambda\right).$$

Define a vector-valued function $F : \mathscr{Z}^m \to \mathcal{H}_K$ by

$$\begin{aligned} F(\mathbf{z}) &= \frac{1}{m^2} S_{\mathbf{x}}^T D Y - \frac{1}{m^2} S_{\mathbf{x}}^T D S_{\mathbf{x}} f_\lambda \\ &= \frac{2}{m^2} \sum_{i,j=1}^m y_i (K_{x_i} - K_{x_j}) + \frac{2}{m^2} \sum_{i,j=1}^m f_\lambda(x_i)(K_{x_i} - K_{x_j}). \end{aligned}$$

Then we can verify that for $i \in \{1, \ldots, m\}$,

$$\|F(\mathbf{z}) - \mathbb{E}_{z_i} F(\mathbf{z})\|_{\mathcal{H}_K} \leq \frac{4\kappa}{m} \{M + \|f_\lambda\|_\infty\}.$$

By independence, we can verify easily that

$$\mathbb{E}_{\mathbf{x}} \frac{1}{m^2} S_{\mathbf{x}}^T D S_{\mathbf{x}} = \frac{m-1}{m} L_K \quad \text{and} \quad \mathbb{E}_{\mathbf{z}} \frac{1}{m^2} S_{\mathbf{x}}^T D Y = \frac{m-1}{m} L_K f_\rho.$$

It follows that

$$\mathbb{E}_{\mathbf{z}} F(\mathbf{z}) = \frac{m-1}{m} L_K (f_\rho - f_\lambda).$$

Since $(L_K + \frac{\lambda}{2} I) f_\lambda = L_K f_\rho$, we have $\frac{\lambda}{2} f_\lambda = L_K(f_\rho - f_\lambda) = \frac{m}{m-1}\mathbb{E}_{\mathbf{z}} F(\mathbf{z})$. Thus,

$$\begin{aligned}\|f_{\mathbf{z},\lambda} - f_\lambda\|_{\mathcal{H}_K} &\le \frac{1}{\lambda}\left\|F(\mathbf{z}) - \frac{\lambda}{2} f_\lambda\right\|_{\mathcal{H}_K} = \frac{1}{\lambda}\left\|F(\mathbf{z}) - \frac{m}{m-1}\mathbb{E}_{\mathbf{z}} F(\mathbf{z})\right\|_{\mathcal{H}_K} \\ &\le \frac{1}{\lambda}\|F(\mathbf{z}) - \mathbb{E}_{\mathbf{z}} F(\mathbf{z})\|_{\mathcal{H}_K} + \frac{1}{2m}\|f_\lambda\|_{\mathcal{H}_K}.\end{aligned}$$

Here the first inequality is derived from Lemma 1.

According to the definition (2), we have

$$\lambda\|f_\lambda\|_{\mathcal{H}_K}^2 \le \mathcal{E}(f_\lambda) + \lambda\|f_\lambda\|_{\mathcal{H}_K}^2 \le \mathcal{E}(0) \le M^2.$$

Then, $\|f_\lambda\|_{\mathcal{H}_K} \le M\lambda^{-\frac{1}{2}}$. Applying Lemma 3 with $\tilde{M} = \frac{4M\kappa(1+\kappa\lambda^{-\frac{1}{2}})}{m}$ and $\sigma^2 \le m\tilde{M}^2$, we derive the desired result. □

To get the total error estimate of $\inf_{f\in\mathcal{G}}\|f_{\mathbf{z},\lambda} - f\|_{\mathcal{H}_K}$, we need to bound the approximation error $\inf_{f\in\mathcal{G}}\|f_\lambda - f\|_{\mathcal{H}_K}$. The approximation error depends on the characteristics of $\mathcal{H}_K$ and ρ. We shall bound it by a functional analysis approach inspired from approximation estimates in [11].

Proposition 5. *Denote as L_K^r the rth power of L_K. If $L_K^{-r} f_\rho \in \mathcal{H}_K$, we have, for $0 < r \le 1$,*

$$\inf_{f\in\mathcal{G}}\|f_\lambda - f\|_{\mathcal{H}_K} \le \|L_K^{-r} f_\rho\|_{\mathcal{H}_K} 2^{-r}\lambda^r.$$

Proof. Define $\tilde{\lambda} = \frac{\lambda}{2}$; we have

$$f_\lambda - f_\rho = (L_K + \tilde{\lambda} I)^{-1} L_K f_\rho - f_\rho = -\tilde{\lambda}(L_K + \tilde{\lambda} I)^{-1} L_K^r L_K^{-r} f_\rho.$$

It follows that

$$\begin{aligned}\|f_\lambda - f_\rho\|_{\mathcal{H}_K} &= \tilde{\lambda}\|(L_K + \tilde{\lambda} I)^{-1} L_K^r L_K^{-r} f_\rho\|_{\mathcal{H}_K} \\ &\le \tilde{\lambda}\|(L_K + \tilde{\lambda} I)^{-1} L_K^r\|\,\|L_K^{-r} f_\rho\|_{\mathcal{H}_K}.\end{aligned} \tag{5}$$

For the function $\varphi(u) = \frac{u^r}{\tilde{\lambda}+u}$, $0 \le r \le 1$, we can check that its maximum is obtained at $u_0 = \frac{r\tilde{\lambda}}{1-r}$. Since L_K is a positive operator from $\mathcal{H}_K$ to $\mathcal{H}_K$, we have

$$\|(L_K + \tilde{\lambda} I)^{-1} L_K^r\| \le \|\varphi\|_\infty = \frac{\left(\frac{r\tilde{\lambda}}{1-r}\right)^r}{\frac{r\tilde{\lambda}}{1-r} + \tilde{\lambda}} = r^r(1-r)^{1-r}\tilde{\lambda}^{r-1} \le \tilde{\lambda}^{r-1}. \tag{6}$$

Combining (5) and (6), we derive the desired statement. □

Theorem 6. *Assume that $L_K^{-r} f_\rho \in \mathcal{H}_K$ for some $0 < r \le 1$. Taking $\lambda = m^{-\frac{1}{2r+3}}$, for any $0 < \delta < 1$, we have, with confidence $1 - \delta$,*

$$\inf_{f\in\mathcal{G}}\|f_{\mathbf{z},\lambda} - f\|_{\mathcal{H}_K} \le C m^{-\frac{r}{2r+3}}\log\frac{2}{\delta}, \tag{7}$$

where C is a constant independent of m, δ.

Proof. Combining Proposition 4 with Proposition 5, we find that, with confidence $1-\delta$, the total error satisfies

$$\inf_{f\in\mathcal{G}}\|f_{\mathbf{z},\lambda}-f\|_{\mathcal{H}_K}\le\|f_{\mathbf{z},\lambda}-f_\lambda\|_{\mathcal{H}_K}+\|f_\lambda-f_\rho\|_{\mathcal{H}_K}\le C\log\frac{2}{\delta}\left\{\frac{1}{\lambda^{\frac{3}{2}}\sqrt{m}}+\lambda^r\right\}.$$

The desired convergence rate follows from the choice of the parameter λ. □

Remark 1. The optimal rate derived from Theorem 6 is achieved by $r=1$. In this case, the learning rate (7) can be arbitrarily close to the order $m^{-\frac{1}{5}}$. Furthermore, the norm $\inf_{f\in\mathcal{G}}\|f_{\mathbf{z},\lambda}-f\|_{\mathcal{H}_K}$ cannot be bounded by the excess error $\mathcal{E}(f_{\mathbf{z},\lambda})-\inf_{f\in\mathcal{G}}\mathcal{E}(f)$, and hence our capability independent convergence analysis for the $\mathcal{H}_K$ norm is new in generalization analysis. Our convergence analysis could be further improved by replacing the condition $L_K^{-r}f_\rho\in\mathcal{H}_K$ with $L_K^{-r}f_\rho\in L^2_{\rho_{\mathcal{X}}}$, and we leave this for future study.

Remark 2. Note that

$$\mathcal{E}(f_{\mathbf{z},\lambda})-\inf_{f\in\mathcal{G}}\mathcal{E}(f)\le 2\|f_{\mathbf{z},\lambda}-f_\rho\|^2_{L^2_{\rho_{\mathcal{X}}}}+8M\|f_{\mathbf{z},\lambda}-f_\rho\|_{L^1_{\rho_{\mathcal{X}}}}.$$

On the basis of Theorem 6, we have, with confidence at least $1-\delta$,

$$\mathcal{E}(f_{\mathbf{z},\lambda})-\inf_{f\in\mathcal{G}}\mathcal{E}(f)=O\left(m^{-\frac{r}{2r+3}}\log\frac{2}{\delta}\right).$$

When $r=1$ and $\lambda=m^{-\frac{1}{5}}$, the learning rate from $\mathcal{E}(f_{\mathbf{z},\lambda})$ to $\inf_{f\in\mathcal{G}}\mathcal{E}(f)$ can be arbitrarily close to the order $m^{-\frac{1}{5}}$.

Acknowledgments

The author is indebted to the anonymous reviewers for their careful comments and constructive suggestions. This work was supported partially by the National Natural Science Foundation of China (NSFC) under grant No. 11001092 and the Fundamental Research Funds for the Central Universities (Program No. 2011PY130, 2011QC022).

References

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, D. Roth, Generalization bounds for the area under the ROC curve, J. Mach. Learn. Res. 6 (2005) 393–425.
[2] S. Agarwal, P. Niyogi, Generalization bounds for ranking algorithms via algorithmic stability, J. Mach. Learn. Res. 10 (2009) 441–474.
[3] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[4] S. Clémencon, G. Lugosi, N. Vayatis, Ranking and empirical minimization of U-statistics, Ann. Statist. 36 (2008) 844–874.
[5] D. Cossock, T. Zhang, Statistical analysis of Bayes optimal subset ranking, IEEE Trans. Inf. Theory 54 (11) (2008) 5140–5154.
[6] S. Mukherjee, D.X. Zhou, Learning coordinate covariances via gradients, J. Mach. Learn. Res. 7 (2006) 519–549.
[7] C. Rudin, R.E. Schapire, Margin-based ranking and an equivalence between adaBoost and rankBoost, J. Mach. Learn. Res. 10 (2009) 2193–2232.
[8] C. Rudin, The P-norm push: a simple convex ranking algorithm that concentrates at the top of the list, J. Mach. Learn. Res. 10 (2009) 2233–2271.

[9] S. Smale, D.X. Zhou, Shannon sampling II. Connections to learning theory, Appl. Comput. Harmon. Anal. 19 (2005) 285–302.

[10] S. Smale, D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constr. Approx. 26 (2007) 153–172.

[11] H.W. Sun, Q. Wu, Application of integral operator for regularized least square regression, Math. Comput. Modeling 49 (2009) 276–285.